29

Stereo Vision Based SLAM Issues and Solutions

D.C. Herath, K.R.S. Kodagoda and G. Dissanayake
ARC Centre of Excellence for Autonomous Systems, University of Technology, Sydney
Australia

1. Introduction

[1]Simultaneous Localization and Mapping (SLAM) has been one of the active research areas in robotic research community for the past few years. When a robot is placed in an unknown environment a SLAM solution attempts to build a perfect map of the environment while localising the robot with respect to this map simultaneously. Traditionally SLAM utilised endogenous sensor data in the process. Successful SLAM implementations using laser (Guivant and Nebot, 2002), sonar and radar (Clark and Dissanayake, 1999) can be found in the literature, which prove the possibility of using SLAM for extended periods of time in indoor and outdoor environments with well bounded results.

Recent extensions to the general SLAM problem has looked in to the possibility of using 3-dimensional features and the use of alternative sensors to traditionally used lasers and radars. Cameras are competitive alternatives owing to the low cost and rich information content they provide. Despite the recent developments in camera sensors and computing, there are still formidable challenges to be resolved before successful vision based SLAM implementations are realised in realistic scenarios. Monocular camera based SLAM is widely researched (Davison et al., 2004; Kwok et al., 2005), however, binocular camera based SLAM is mostly overlooked. Some of the noted stereo implementations can be found in (Davison and Murray, 2002) and recently in (Jung, 2004). Lack of enthusiasm for research in this direction could possibly be attributed to the misconception that range and bearing information provided by the stereo vision system is directly utilizable providing a simplistic solution to SLAM which is academically less appealing or the apparent success in single camera SLAM implementations.

However, after rigorous analysis and sensor modelling, we found that the standard extended Kalman filter (EKF) based SLAM with small base line stereo vision systems can easily become inconsistent (Herath et al., 2006a) .

This chapter attempts to provide readers with an understanding of the SLAM problem and its solutions in the context of stereo vision. The chapter introduces the Extended Kalman Filter as applied to the generic SLAM Problem. Then, while identifying the prevailing issues inherent in solutions to the SLAM problem in stereo vision context, our solutions are presented with simulated and experimental evaluations. Several components of the stereo

1 This work is supported by the ARC Centre of Excellence program, funded by the Australian Research Council (ARC) and the New South Wales State Government.

Source: Vision Systems: Applications, ISBN 978-3-902613-01-1
Edited by: Goro Obinata and Ashish Dutta, pp. 608, I-Tech, Vienna, Austria, June 2007

vision system, including outlier rejection, sensor modelling, inconsistency analysis and alternate formulations of SLAM are discussed.

2. Simultaneous Localisation and Mapping (SLAM)

This section presents an introduction to the Kalman filter in the context of Simultaneous Localization and Mapping beginning with the derivation of the standard Kalman filter equations for a linear discrete system and then extending them to accommodate real world non linear systems, the Extended Kalman Filter (EKF) as implemented in majority of the SLAM solutions.

2.1 Linear Discrete-Time Kalman Filter

In order to derive the Kalman filter for discrete linear system, its process and observation models must be defined. The Kalman Filter consists of three recursive stages. (1) Prediction, (2) observation and, (3) update Stage. For a linear, discrete-time system the state transition equation (process model) can be written as follows

$$\mathbf{x}(k) = \mathbf{F}(k)\mathbf{x}(k\text{-}1) + \mathbf{B}(k)\mathbf{u}(k) + \mathbf{G}(k)\mathbf{v}(k) \tag{1}$$

Where $\mathbf{x}(k)$ - state at time k, $\mathbf{u}(k)$ - control input vector at time k, $\mathbf{v}(k)$ - additive process noise, $\mathbf{B}(k)$ - control input transition matrix, $\mathbf{G}(k)$ - noise transition matrix and $\mathbf{F}(k)$ - state transition matrix. The linear observation equation can be written as

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{w}(k) \tag{2}$$

where $\mathbf{z}(k)$ - observation made at time k, $\mathbf{x}(k)$ - state at time k, $\mathbf{H}(k)$ - observation model and $\mathbf{w}(k)$ - additive observation noise. Process and observation noise are assumed to be zero-mean and independent. Thus

$$E[v(k)]=E[w(k)] = 0, \forall k \text{ and } E[v_i w_j^T] = 0, \forall i, j$$

Motion noise and the observation noise will have the following corresponding covariance;

$$E[v_i v_j^{\mathrm{T}}] = \delta_{ij}\mathbf{Q}_i, \quad E[w_i w_j^T] = \delta_{ij}\mathbf{R}_i$$

The estimate of the state at a time k given all information up to time k is written as $\hat{\mathbf{x}}(k/k)$ and the estimate of the state at a time k given information up to time k-1 is written as $\hat{\mathbf{x}}(k/k-1)$ and is called the prediction. Thus given the estimate at (k-1) time step the prediction equation for the state at k^{th} time step can be written as

$$\hat{\mathbf{x}}(k/k-1) = \mathbf{F}(k)\hat{\mathbf{x}}(k-1/k-1) + \mathbf{B}(k)\mathbf{u}(k) \tag{3}$$

And the corresponding covariance prediction;

$$\mathbf{P}(k/k-1) = \mathbf{F}(k)\,\mathbf{P}(k-1/k-1)\mathbf{F}^T(k) + \mathbf{G}(k)\mathbf{Q}(k)\mathbf{G}^T(k) \tag{4}$$

Then the unbiased (the conditional expected error between estimate and true state is zero) linear estimate is

$$\hat{\mathbf{x}}(k/k) = \hat{\mathbf{x}}(k/k-1) - \mathbf{W}(k)[\mathbf{z}(k) - \mathbf{H}(k)\hat{\mathbf{x}}(k/k-1)] \tag{5}$$

Where **W**(k) is the Kalman Gain at time step k. This is calculated as:

$$\mathbf{W}(k) = \mathbf{P}(k/k-1)\mathbf{H}^T(k)\mathbf{S}^{-1}(k) \tag{6}$$

Where **S**(k) is called the innovation variance at time step k and given by:

$$\mathbf{S}(k) = \mathbf{H}(k)\mathbf{P}(k/k-1)\ \mathbf{H}^T(k) + \mathbf{R}(k) \tag{7}$$

and the covariance estimate is

$$\mathbf{P}(k/k) = (\mathbf{I} - \mathbf{W}(k)\mathbf{H}(k)\mathbf{P}(k/k-1)(\mathbf{I} - \mathbf{W}(k)\mathbf{H}(k))^T + \mathbf{W}(k)\mathbf{R}(k)\mathbf{W}^T(k) \tag{8}$$

Essentially the Kalman filter takes a weighted average of the prediction $\hat{\mathbf{x}}(k/k-1)$, based on the previous estimate $\hat{\mathbf{x}}(k-1/k-1)$, and a new observation $\mathbf{z}(k)$ to estimate the state of interest $\hat{\mathbf{x}}(k/k)$. This cycle is repeatable.

2.2 The Extended Kalman Filter

Albeit Kalman filter is the optimal minimum mean squared (MMS) error estimator for a linear system, hardly would one find such a system in reality. In fact the systems considered in this chapter are purely non-linear systems. A solution is found in the Extended Kalman Filter (EKF) which uses a linearised approximation to non-linear models. The extended Kalman filter algorithm is very similar to the linear Kalman filter algorithm with the substitutions;

$\mathbf{F}(k) \rightarrow \mathbf{f}_x(k)$ and $\mathbf{H}(k) \rightarrow \mathbf{h}_x(k)$, where $\nabla\mathbf{f}_x(k)$ and $\nabla\mathbf{h}_x(k)$ are non-linear functions of both state and time step, and $\mathbf{f}_x(k)$, $\mathbf{h}_x(k)$ are the process model and observation model respectively. Therefore the main equations in EKF can be summarized as follows;

1. Prediction equations

$$\hat{\mathbf{x}}(k/k-1) = \mathbf{f}(\hat{\mathbf{x}}(k-1/k-1)\ , \mathbf{u}(k)) \tag{9}$$

$$\mathbf{P}(k/k-1) = \nabla\mathbf{f}_x(k)\ P(k-1/k-1)\nabla^T\mathbf{f}_x(k) + \mathbf{Q}(k) \tag{10}$$

2. Update equations

$$\hat{\mathbf{x}}(k/k) = \hat{\mathbf{x}}(k/k-1) + \mathbf{W}(k)[\mathbf{z}(k) - \mathbf{h}(k/k-1)] \tag{11}$$

$$\mathbf{P}(k/k) = \mathbf{P}(k/k-1) - \mathbf{W}(k)\mathbf{S}(k)\mathbf{W}^T(k) \tag{12}$$

Where

$$\mathbf{S}(k) = \nabla\mathbf{h}_x(k)\mathbf{P}(k/k-1)\nabla^T\mathbf{h}_x(k) + \mathbf{R}(k) \tag{13}$$

2.3 Filter Consistency

The SLAM formulation presented in the previous section represents the posterior as a unimodal Gaussian. Thus the state estimates are parameterized by what is known as the *moments parameterization*. An important ramification of this representation is that not only it represents the current mean $\hat{\mathbf{x}}(k/k)$ but also gives an estimate of the covariance $\hat{\mathbf{P}}(k/k)$, and when the filter is *consistent,* the estimated covariance should match the Mean Square Error of the true distribution. As will be discussed in the following section this is widely used in interpreting EKF based SLAM results.

However a more appropriate measure of consistency when the true state $\mathbf{x}_k$ is known could be arrived at using the normalized estimation error squared (NEES) as defined by (Bar-Shalom et al., 2001),

$$\varepsilon(k) = (\mathbf{x}(k) - \hat{\mathbf{x}}(k/k))^T \mathbf{P}(k/k)^{-1} (\mathbf{x}(k) - \hat{\mathbf{x}}(k/k)) \tag{14}$$

Under the hypothesis that filter is consistent and is linear Gaussian, ε_k is chi-square distributed with n_x degrees of freedom. Where n_x is the dimension of $\mathbf{x}_k$.

$$E[\varepsilon(k)] = n_x \tag{15}$$

Using multiple Monte Carlo simulations to generate N independent samples, the average NEES can be calculated as

$$\bar{\varepsilon}_k = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{ik} \tag{16}$$

Then under the previous hypothesis $N\bar{\varepsilon}(k)$ will have a chi-square density with Nn_x degrees of freedom. Then the above hypothesis is accepted if

$$\bar{\varepsilon}(k) \in [r_1, r_2] \tag{17}$$

where the *acceptance interval* is determined on a statistical basis.

2.4 An Example

To illustrate the formulation of the standard EKF, lets consider an example where a simple differential driven robot traversing on a 2D plane. The robot is equipped with a sensor capable of making 3D measurements to point features in the environment (Fig. 1). The robot state is defined by $X_r = [x_r \quad y_r \quad \varphi_r]^T$, where x_r and y_r denotes location of the robot's rear axle centre with respect to a global coordinate frame and φ_r is the heading with reference to the x-axis of the same coordinate system. Landmarks are modelled as point features, $\mathbf{p}_i = [x_i \quad y_i \quad z_i]^T$, i = 1,…,n. The vehicle motion through the environment is modelled as a conventional discrete time process model as in (9).

$$\begin{bmatrix} x_r(k+1) \\ y_r(k+1) \\ \varphi_r(k+1) \end{bmatrix} = \begin{bmatrix} x_r(k) + \Delta T V(k) \cos(\varphi_r(k)) \\ y_r(k) + \Delta T V(k) \sin(\varphi_r(k)) \\ \varphi_r(k) + \Delta T \omega(k) \end{bmatrix} \tag{18}$$

ΔT is the time step, V(k) is the instantaneous velocity and $\omega(k)$ is the instantaneous turn-rate. The observation model can be represented as,

$$Z\ (k+1) = \begin{bmatrix} z_x(k+1) \\ z_y(k+1) \\ z_z(k+1) \end{bmatrix} = \begin{bmatrix} a \\ b \\ z_{fi}(k+1) \end{bmatrix} \tag{19}$$

where

$$a = (x_{fi}(k+1) - x_r(k+1))\cos(\varphi_r(k)) + (y_{fi}(k+1) - y_r(k+1))\sin(\varphi_r(k))$$

$$b = -(x_{fi}(k+1) - x_r(k+1))\sin(\varphi_r(k)) + (y_{fi}(k+1) - y_r(k+1))\cos(\varphi_r(k))$$

It is to be noted that each feature is defined by a point in 3D space, $\mathbf{x}_{fi}(k) = [x_{fi}(k), y_{fi}(k), z_{fi}(k)]^T$.

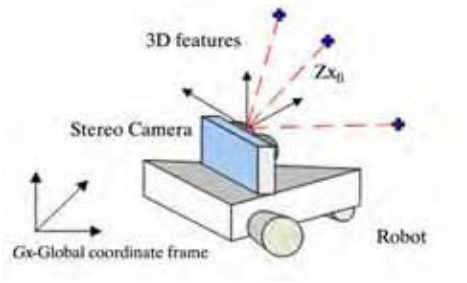
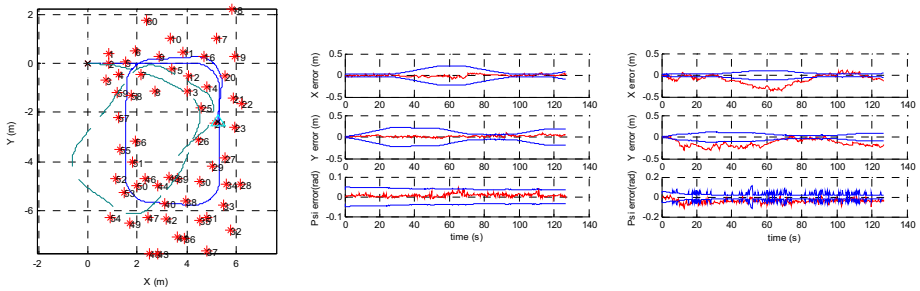


Figure 1. The robot in 3D world coordinates observing a feature in 3D space.

Fig. 2 (a) shows a simulated environment with the path robot has taken amongst the 3D features. Fig. 2 (b) depicts the results of this example implementation on the simulated environment. The three graphs depict the three components of the robot pose. In the top graph of Fig. 2 (b), the middle line represents error between the EKF estimate and the actual value of the x-component of the robot pose against the time. The two outer lines mirroring each other are the 2-standard deviation estimates (2-*sigma*). When the filter is *well tuned* the error lies appropriately bounded within these 2-*sigma* limits. Fig. 2 (c) illustrates a case of filter inconsistency where the filter has become optimistic.



a.) b.) c.)

Figure 2. (a) Simulated environment: solid line – true path, dashed line – odometry path, * - features. (b) State errors with estimated 2-sigma bounds for a well tuned filter (c) An inconsistent filter

3. Stereo Vision

Generally, more precise the sensors used in SLAM more tractable and practical the solution is. Underlying characteristics of the sensor play an important role in determining the scale and practical use of the SLAM algorithm. Sensors such as laser have proven to be very precise in nature and have shown to work well in large environments for extended periods of time (Guivant, 2002; 2003; Wang, 2004). However vision is yet to prove its application in similar environments. In vision, successful implementations to date have used either large baseline stereo cameras (Davison, 1998; Jung, 2004), camera configurations with more than two cameras(Se et al., 2002) providing refined observations or single camera bearing only (Kwok and Dissanayake, 2003; 2004) methods. Principal aim of this section is to asses the performance of a small baseline binocular stereo camera equipped with wide angle lenses in the context of robotic SLAM.

3.1 The Sensor

Stereopsis or Stereoscopic vision is the process of perceiving depth or distances to objects in the environment. As a strand of computer vision research stereo vision algorithms have advanced noticeably in the past few decades to a point where semi-commercial products are available as *off the shelf* devices. However a more augmented approach is needed to realize a sensor useful in SLAM. Following list is an attempt to enumerate the essential components of such a sensor in the context of SLAM.
(1.) Stereo camera-hardware for acquiring stereo images (2.) Calibration information-contains intrinsic and extrinsic information about the camera necessary for correcting image distortion and depth calculation (3.) Interest point (features) selection algorithm-mechanism through witch naturally occurring features in the environment are selected for integration in the state vector (4.) Feature tracking algorithm-Image based mechanism used for data association (5.) Stereo correspondence algorithm-estimates the disparity at corresponding pixels (6.) Filtering-mechanisms used to remove spurious data. A schematic of the components along with interactions amongst each other is outlined in Fig. 3.

3.2 Sensor Error Analysis

As mentioned in the beginning of the chapter characteristics of a sensor dictates the limits of its applications. In the following sections a discussion of an empirical study of the particular sensor of interest is given based on two representative experiments conducted. It was found that even though quantitative error analyses of stereo, based on static cameras are available in the literature they do not necessarily represent the effects of a moving camera. This study fills a void on specific characterisation of noise performance of small baseline large field of view camera in respect to SLAM. In this context several robotic mapping experiments were carried out in order to understand the behaviour of sensor noise.

From previous section on camera modelling the triplet $\mathbf{z} = [u, v, d]^T$ forms the principal observation $\mathbf{z}$ by the sensor. Where (u,v) being the image coordinates of a feature and d is the disparity. Assuming that the errors in the observations to be additive $\mathbf{z}$ can be written as,

$$\mathbf{z} = \mathbf{z}_{true} + \mathbf{v}(\zeta, \mathbf{z}_{true}) \tag{20}$$

Where $\mathbf{z}_{true}$ being the true state of the observation and $\mathbf{v}$ being the additive noise component dependent on the sensor characteristics ξ and on the true state itself as will be shown empirically later. Modelling and understanding the behaviour of $\mathbf{v}$ is the subject of discussion in sections 3.4 and 3.5. In section 3.6 the discussion continues on modelling the error behaviour of the projected form $\mathbf{z}_c$ of this observation in to the 3D coordinate frame.

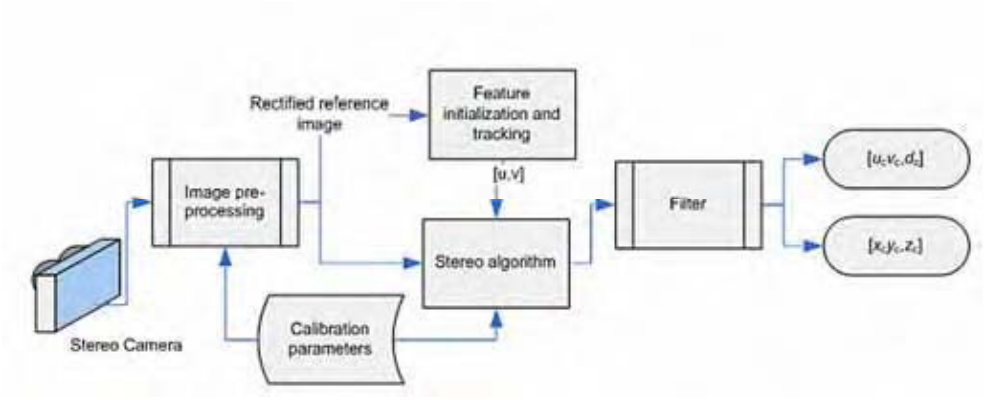
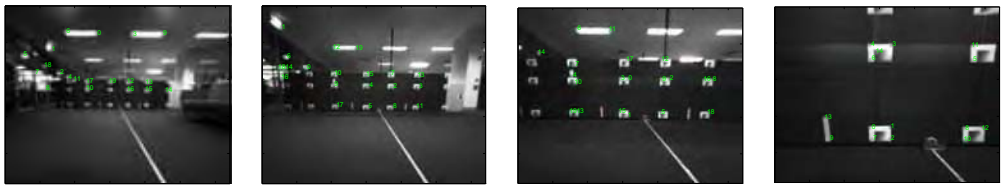


Figure 3. The vision system for a SLAM implementation

3.3 Mapping Experiments

References are made to the two experiments described below in the following sections.
Experiment 1- A pioneer robot mounted with the stereo camera was moved on a controlled path while capturing set of images at each 0.02m interval. The feature selection algorithm was allowed to select 30 features at the beginning of the sequence. The tracking algorithm attempts to track these features between consecutive images.
Experiment 2- Again the robot was moved on a controlled path while observing artificial features laid on a large vertical planar surface. Features were laid out so as to cover the whole field of view of the cameras. A SICK laser was used to maintain parallel alignment between the camera and the surface and to measure the nominal distance between the robot and the surface. Robot was moved in 0.05m increments from a distance of 6m to 1m.
In this experiment 20 features were initialised at each stop and were then tracked for 29 consecutive images. For analysis of this data, at least 9 features were selected manually covering the widest possible area of the planar surface at each stop point. This set of features would then represent the expected sensor behaviour at the given distance.



a.) b.) c.) d.)

Figure 4. Rectified images overlaid with features at (a) 5.4 m (b) 3.9 m (c) 2.4 m (d) 0.9 m

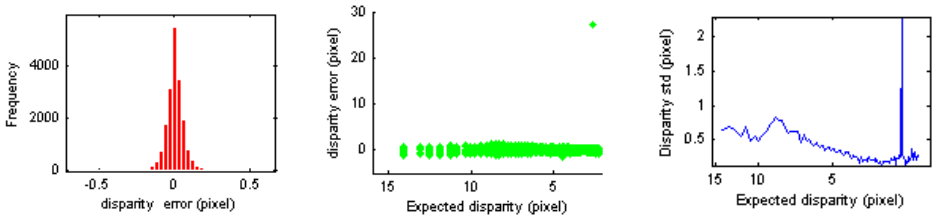
3.4 Uncertainty in Disparity

In order to establish an error model for the disparity an analyses based on the finite data series from experiment 2 was performed. The data presents a unique perspective on the variance in disparity as observations are made at varying distances. In this case an approximate range between 1m and 6m inclusively. This depth range translates to an effective disparity range approximately between 1 and 15 pixels. The stereo correlation algorithm is set to search for a pixel range between 1 and 32.

Following general statistical procedures it is possible to estimate a set of parameters that represent the disparity observation process based on this finite sequence of data. Fig. 5 (a) shows the overall variation in disparity. This is based on the calculated disparity at each individual feature that were manually selected in each initial image combined with all the points that were tracked consecutively are pooled together by subtracting the disparity means corresponding to each individual tracking sequence.

Although by the analysis of the autocorrelation it is easily established that the process is 'white' the general assumption of the distribution being Gaussian is an oversimplification of the true distribution. Especially in the case of small baseline cameras and wide-angle lenses this variation is a complex combination of local biases introduced in the image rectification process and stereo correlation mismatches undetected by the various filtering mechanisms. The distortions introduced by wide-angle lenses induce biases at each pixel in the image. Even though they are constant it is extremely difficult to accurately measure the individual component at pixel level. Also the area correlation algorithm used to estimate the disparities itself is prone to gross errors depending on the construct of the environment in which the images are captured.

In order to understand these subtle variations it is best to analyse the variation in disparity at different depths independently. Fig. 5 (b) shows the variation in observed disparity against the expected disparity. Again the data from experiment 2 are used in the analysis. In this the disparities of the features selected at each distance along with the consecutively tracked points are pooled together and the resulting combined data are subtracted from the population mean.



a.) b.) c.)

Figure 5. Disparity error. (a) Distribution (b) Zero mean error distribution with depth (c) Zero mean standard deviation (log scale). The spike in standard deviation is due to a stereo mismatch that was not detected by any of the heuristics applied in stereo correspondence algorithm

Several observations can be made. Firstly, data still contains many visible outliers that are difficult to be eliminated by the various smoothing operations. Secondly a rather intuitive observation is the correlation in the variance of the disparity distribution with the expected disparity. As would be expected variance is smaller for features seen from afar and it increases gradually with nearby features. For faraway features the disparity is small and also the discriminatory information contained within the correlation area is higher compared to a closer observation. This is especially true for environments where lack of texture persists. This gives a higher confidence to the disparity values estimated for features afar as opposed to ones closer. This is a better interpretation for the variance in disparity and based on this interpretation it is better to assume a varying disparity standard deviation correlated with the estimated disparity value as opposed to the general practice of assuming a constant disparity standard deviation. The observation standard deviation is shown in Fig. 5 (c).

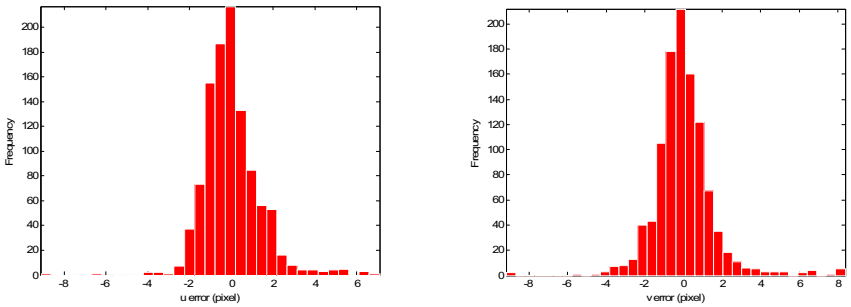
It is difficult to estimate an exact relationship between the disparity variation with the estimated disparity. Thus an empirically generated curve based on the results shown in Fig. 5 (b) is used. It was also observed that the variance estimated thus is slightly higher than the one shown above in Fig. 5 (a). This stems from the fact that the local biases are present in the data shown in Fig. 5 (b). This can be illustrated by scrutinising the local distributions present in the disparity data corresponding to each feature location at a given depth. Fig. 8 shows an example local distributions contributing to the overall distribution at a given depth. As can be noticed there are independent local distributions dispersed from the true expected mean. These are a combination of local biases in the image, stereo mismatches and any misalignments of the stereo hardware and the reference system. For practical purposes correcting these errors is difficult and an all encompassing error model is thus adapted.

3.3 Uncertainty in *u* and *v*

In order to model the errors in u and v for SLAM a dynamic camera error model needs to be studied which would include the behaviour of the tracking algorithm as well as other dynamics involved with the camera motion. From experiment 1 and 2 it is possible to extract a representative set of data for this purpose. Again as discussed for the case of disparity error, u and v also carries components of local bias due to distortion effects and other misalignments. In addition the effects of the feature tracker also contribute when the augmented sensor representation is considered.

For this analysis only a single image is considered at each depth. These images are then assembled from a depth of 1m to 6m. 16 features covering the entire image plane are then initialised in the image corresponding to 1m depth and are then consecutively tracked through to image at 6m depth. This while tracking a set of features at fixed locations in space will map to varying u,v coordinates. This essentially captures the overall behaviour of u,v in the entire image plane.

Fig. 6 shows the results for both parameters where cumulative data for each point is subtracted by the expected values at each point and then combined together. Qualitatively these results resemble Gaussian distributions. However it is possible to observe various artefacts appearing in the tails of the distributions indicating that a considerable amount of spurious data is present for the reasons discussed earlier. This spuriousness in u,v and d pose considerable challenges to a successful implementation of a SLAM algorithm. Various issues arising from these observations are discussed in the next section



a.) b.)

Figure 6. Error distribution (a) in u with standard deviation = 1.34 (b) in v with standard deviation = 1.53

4. Issues and Solutions

Primary goal of this chapter is to elucidate several theoretical and practical issues that have been noted during many implementations of stereo vision based SLAM. In this section a series of such issues that contributes to filter divergence, increase in computational burden and/or complete failure of the filter are illustrated. In each sub-section an issue is presented first with its effects on the algorithm and then possible solutions in averting the consequences are discussed.

4.1 Limited Field of View

One of the most fundamental issues that plague vision based SLAM is the limited field of view (FOV) of the sensor. When compared to traditional sensors like laser and radar the FOV of vision sensors are 20~40% narrower. Even though the 2D structure of the sensor affords more information the narrow FOV limits the ability to observe features for prolonged periods, a desirable requirement to reduce error bounds in the state estimations – a corollary of the results proven in (Dissanayake et al., 2001). As noted in several works(Bailey et al., 2006; Huang and Dissanayake, 2006), notably the increase in heading uncertainty tends to increase the possibility of filter divergence. This has been observed in our implementations, especially in confined office like environments where many of the features observed vanishing from the FOV rapidly and re-observation of them delayed until a large loop is closed.

Slight improvement to this situation is brought through the introduction of wide angle lenses. However, the choice is a compromise between the sensor accuracy and the FOV. Wide angle lenses suffer from noticeable lens distortion (Fig. 7 (a)) and the rectification (Fig. 7 (b)) process introduces errors. One undesirable effect of using such lenses is the local biases in disparity calculation. To illustrate this consider a static camera observing a perfect plane which is parallel to the camera x-y plane. Disparity results of observing several features on this plane are plotted in Fig. 8. As can be seen the biases at various points are noticeable and are high as 2-pixels. This at most violates the fundamental assumption of Gaussian noise model in SLAM.In order to alleviate this issue it is necessary to estimate and apply radial distortion parameters in the rectification process. Also in severe cases, or when higher accuracy is demanded look-up tables are suggested.

a.) b.)

Figure 7.An image from a wide angle lens. a.) raw. b.) rectified

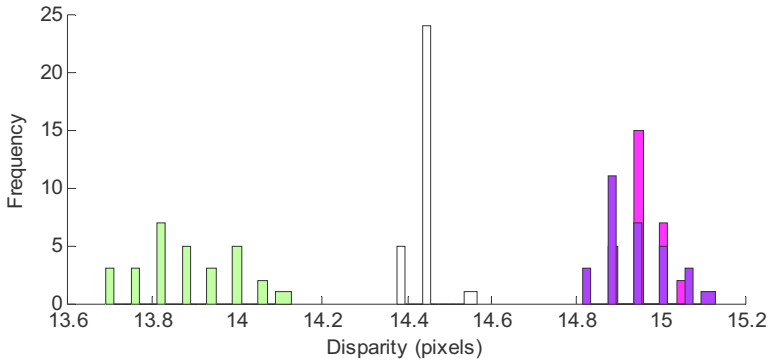


Figure 8. Local biases in disparity (Expected mean disparity is 14.7)

4.2 Number of Features

Each feature added to the filter contributes new information. However with each new feature added to the state vector increases the computational burden. Even with sophisticated algorithms available computational complexity of SLAM still remains high and grows with each added state. Also depending on the data association mechanism used the ambiguity of features could increase leading to false association. This necessitates a reliable ranking mechanism (Shi and Tomasi, 1994) to optimize the number of features processed per image. The ranking criteria should not only look at which are "good features to track" but also its viability as a 3-D observation. Therefore it is possible to integrate other stereo confidence measures like uniqueness in to the ranking mechanism. Such an integrated approach alleviates selecting features that are ineffective as 3D measurements.

Another common issue seen especially in indoors with highly structured built environments is the lapses in suitably textured surfaces needed to generate reliable features and depth measurements. In extreme cases we have observed heavy reliance on other sensors such as odometry in filters. This is a limitation on point feature based implementations and alternative feature forms such as lines and curves would be more appropriate depending on the environment in which the application operates.

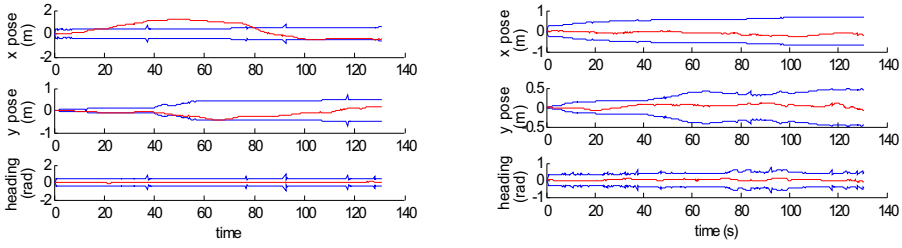
The minimum number of features per image is also dictated partially by the environment the application operates as well as the accuracy of the stereo algorithm. As shown earlier the depth accuracy correlates with the depth measured. Thus it is necessary to observe both features that are closer to the camera for short term translational accuracy as well as ones that are further away for long term rotational accuracy. An issue with most feature selectors is that they tend to cluster around small patches of highly textured areas in a scene. This may or may not result in satisfying the condition stated above. In our experience the best value for minimal number of features is thus selected by repeated experimentations in the intended environment.

4.3 Spurious Features

Spurious features occur not only due to structure (e.g. Occlusion) but also due to gross errors in stereo calculations. For instance in Fig. 9 (a) the pole marked with the arrow and the horizontal edge of the partition in the foreground are two distinct disjoint enteritis. However on the image plane the apparent intersection of the two edges is a positive feature location. Such occlusions results in physically non existent features. These features are catastrophic in a SLAM implementation. A possible method was discussed in (Shi and Tomasi, 1994) in identifying such occlusions by a measure of *dissimilarity*.

a.) b.) c.)

Figure 9. Spurious observations. (a) A rectified image showing several edge profiles. (b) Disparity image (c) Close-up view of the depth profile with a mismatch (see discussion for details)



a.) b.)

Figure 10. Robot pose error with 2-*sigma* error bounds (a) effects of spurious data (b.) with the RanSaC like filter applied

Depending on the image composition it is possible to generate occasional mismatches (Fig. 9(c)) in stereo correspondence. Most stereo algorithms include multiple heuristics (Konolige, 1997) to alleviate this issue. However it is still advisable to include a statistical validation gate (Cox, 1993) for the occasional mismatch that is not filtered by such heuristics.

A third set of spurious features were observed due to feature tracking mechanisms used. These features tend to drift arbitrarily in the image plane. Such features not only are harder to detect by conventional statistical validation gates but also tend to contribute to filter inconsistencies. A solution to such spurious features based on the RanSaC (Fischler and Bolles, 1981) algorithm was discussed in (Herath et al., 2006b). Fig. 10 shows SLAM results for a real data set with (Fig. 10 (b)) and without (Fig. 10 (a)) the RanSaC like filter while maintaining other filter parameters identical. In this instance the consistency has improved, however, inflated observation noise parameters are used in both cases to accommodate the nonlinearities (see 4.5) in the observation model.

4.4 Static vs. Dynamic Noise Parameters

Most researchers tend to use static noise parameters in their SLAM implementations. These are the noise parameters obtained by observing static features through a static camera. However a more realistic set of values can be obtained by estimating these parameters through data obtained by a moving camera especially in the same application environment. An experiment of this nature was discussed in (Herath et al., 2006a). This encompasses not only the error variation in camera, but also the error variations in the feature tracker and other difficult to quantify dynamic factors. This invariably tends to increase the stereo noise parameters and in some cases is much higher than the theoretical sub-pixel accuracies quoted by stereo algorithms.

Another aspect of noise parameters was illustrated in section 3.2. For a better estimate of the noise parameters it is possible to utilise the empirical knowledge of variation in disparity standard deviation with measured depth. Also in (Jung and Lacroix, 2003) presented another observation, where the variation in disparity standard deviation is correlated with the curvature of the similarity score curve at its peak. This knowledge can enhance the quality of the estimation process.

4.5 Nonlinearity Issues

Realistic SLAM problems are inherently non linear. While EKF implementations are shown to be able to handle this nonlinearity an emerging debate in recent years suggest that the nonlinearity could lead to filter inconsistency(Bailey et al., 2006; Huang and Dissanayake, 2006; Julier and Uhlmann, 2001).

These studies concentrate on eventual failure of the filter in large scale and/or long term SLAM implementations. On the other hand the few stereo vision based EKF solutions present in the literature altogether neglects the filter consistency analysis. It is well known that the standard geometric projection equations used in stereo vision are highly nonlinear and suffers from inherent bias (Sibley et al., 2006; 2005). It is imperative then that an analysis is carried out to estimate the effects of this nonlinearity in the context of EKF SLAM. For this reason a set of Monte Carlo simulations were conducted and were analysed using the NEES criterion presented in section 2.3. The simulated environment presented in section 2.4 (Fig. 2 (a)) was used throughout these Monte Carlo runs. N [=50] runs were carried out for each implementation with [2.36, 3.72] being the 95% *probability concentration region* for $\bar{\varepsilon}(k)$ since the dimensionality of the robot pose is 3.

In Fig. 11 (a) the average NEES for the example in 2.4 is shown to be well bounded. This indicates that for the small loop considered in this example a standard EKF yields consistent results. For this simulation, the observation noise ($\mathbf{R}(k)$) has components

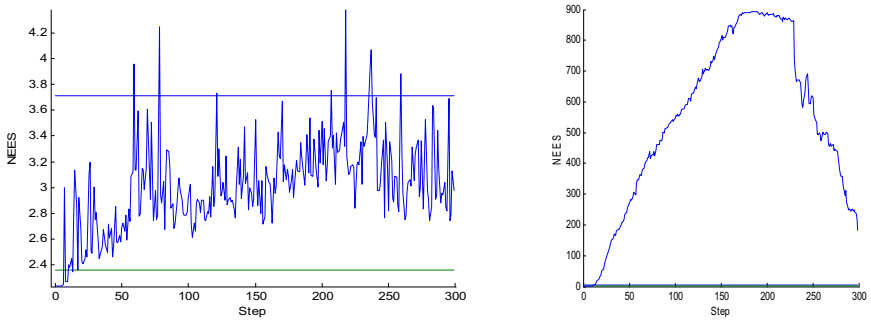
(σ_x=σ_y=σ_z=0.05m) and process noise (**Q**(*k*)) will remain at (σ_v=0.05m/s,σ_w=5deg/s) for all the simulations.

In the second simulation while adhering to the previous formulation, the observations are now subjected to the geometric transformations of a standard stereo vision sensor.

$$x = \frac{Bf}{d}; y = \frac{-Bu}{d}; z = \frac{-Bv}{d} \tag{21}$$

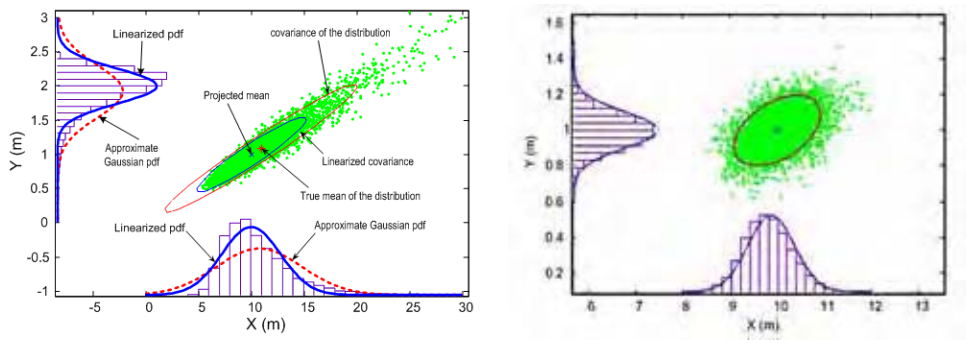
Where *B* is the camera baseline and *f* the focal length. As discussed in the previous section Gaussian noise can be assumed for (*u*,*v*,*d*) and a transformed noise matrix must be used (Herath et al., 2006a) for **R**(*k*). For all the simulations following noise values (σ_u=1.34,σ_v=1.53,σ_d=0.65) estimated from experimental analysis were used. The average NEES results for this simulation are presented in Fig. 11 (b). The unacceptably large values for the statistics indicate that a straight forward SLAM implementation does not yield consistent results. An important parameter in this experiment is the small baseline (*B*) used. At a nominal 9cm this corresponds to a commercially available stereo head on which most of our real experiments are based on. It is possible to show through simulation that lager baselines give rise to lower nonlinearity effects. However it remains a key factor for most stereo heads used in indoor and outdoor scenarios.

To further illustrate this phenomenon, consider the Gaussian random variable $[d,u]^T$ (only two components used for clarity) representing the disparity and horizontal image coordinate for a given feature at $x_c = 10m$ and $y_c = 1m$. With B =0.09m and f = 150 pixels, this translates to mean disparity, d of 1.32 pixels and mean u of 15 pixels. A Monte Carlo simulation can be carried out using (21) to transform Gaussian distributed $[d,u]^T$ into $\left[z_x, z_y\right]^T$. Fig. 12 (a) and (b) show the resulting distributions with 0.09m and 0.5m as baselines respectively. This clearly indicates the non Gaussian nature of the transformed observations when a small baseline camera is used (Fig. 12 (a)). The smaller the baseline is the shorter the range is at which the nonlinear effect manifest.



a) b)

Figure 11. Average NEES of the robot pose (a) Standard EKF (b) Standard EKF with stereo observations



a.) b.)

Figure 12. Errors in projective mapping (a) B=0.09m (b) B=0.5m, linearized and approximated Gaussians are overlapping

A different choice of observation model is tested that yields improved results. As shown above the main cause for the inconsistency is due to the highly nonlinear projective mapping. It is possible to derive a formulation where the principal observation becomes (u,v,d) instead of the widely used (x,y,z) as follows. (Compare this with (19))

$$\hat{\mathbf{z}}(k+1) = \begin{bmatrix} \hat{z}_u(k+1) \\ \hat{z}_v(k+1) \\ \hat{z}_d(k+1) \end{bmatrix} = \frac{f}{x}\begin{bmatrix} -y & -z & B \end{bmatrix}^T \tag{22}$$

where

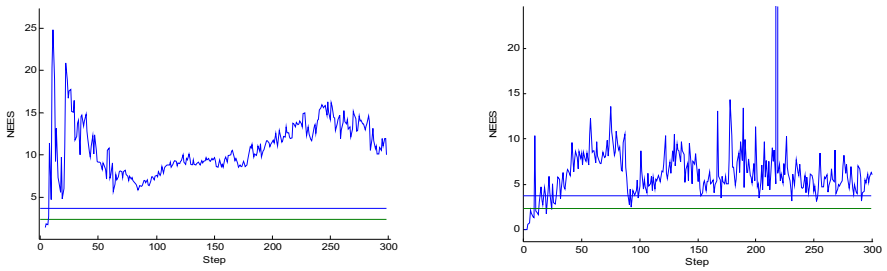
$$x = (\hat{x}_{fi}(k+1) - \hat{x}_r(k+1))\cos(\phi(k)) + (\hat{y}_{fi}(k+1) - \hat{y}_r(k+1))\sin(\phi(k))$$

$$y = -(\hat{x}_{fi}(k+1) - \hat{x}_r(k+1))\sin(\phi(k)) + (\hat{y}_{fi}(k+1) - \hat{y}_r(k+1))\cos(\phi(k))$$

$$z = \hat{z}_{fi}(k+1)$$

This alleviates necessity of the linearized transformation of the noise matrix ($\mathbf{R}(k)$) as measurements are well represented with Gaussian models. Simulation results with the new observation model for average NEES are presented in Fig. 13 (a). Although the improvement over previous model is apparent, filter still remains optimistic. Finally the unscented Kalman filter (UKF) (Julier and Uhlmann, 2004) is implemented with the previous observation model. The UKF performs a *derivative free* transform of the states resulting in better estimates. UKF is shown to work well with highly non linear systems. However the Monte Carlo simulation results indicate (Fig. 13 (b)) that the improvement against consistency is minimal.

These observations lead us to the belief that standard SLAM implementations could yield inconsistent results even for comparatively smaller loops given small baseline stereo cameras are used. An observation hitherto has not been studied. Current solutions for this issue remains at either in use of wider baseline cameras or in the implementation of small loops with sub map (Williams, 2001) like ideas. Better consistency could also be expected by improving the overall noise performance of the vision system. This includes improving the stereo correspondence, resolution of the images as well as improving the stability of the mobile platform.



a) b

Figure 13. Average NEES of the robot pose (a) (*uvd*)-observation model (b) UKF

5. Conclusion

In this chapter we have made an attempt to analyse the issues in stereo vision based SLAM and proposed plausible solutions. Correct sensor modelling is vital in any SLAM implementation. Therefore, we have analyzed the stereo vision sensor behaviour experimentally to understand the noise characteristics and statistics. It was verified that the stereo observations in its natural form (i.e. [*u,v,d*]) can safely be assumed to represent Gaussian distributions. Then several SLAM implementation strategies were discussed using stereo vision. Issues related to limited field of view of the sensor, number of features, spurious features, noise parameters and nonlinearity in the observation model were discussed. It was shown that the filter inconsistency is mainly due to inherent nonlinearity presence in the small baseline stereo vision sensor. Since UKF is more capable in handling nonlinearity issues than that of EKF, an UKF SLAM implementation was tested against inconsistency. However, it too leads to inconsistencies. This shows that even with implementations that circumvent the critical linearization mechanism in standard EKF SLAM as in UKF, the nonlinearity issue in the stereo vision based SLAM can not be resolved. In order to address the filter inconsistency a more elegant solution is currently being researched based on smoothing algorithms which shows promising results.
In conclusion this chapter dwelt on some obscure issues pertaining to stereo vision SLAM and work being done in solving such issues.

6. References

Bailey, Tim, Juan Nieto, Jose Guivant, Michael Stevens and Eduardo Nebot. (2006). Consistency of the EKF-SLAM Algorithm. In I*nternational Conference on Intelligent Robots and Systems* (IROS 2006). Beijing, China.

Bar-Shalom, Yaakow, X.-Rong Li and Thiagalingam Kirubarajan. (2001). *Estimation with Applications to Tracking and Navigation*. Somerset, New Jersey: Wiley InterScience.

Clark, S. and G. Dissanayake. (1999). Simultaneous localisation and map building using millimetre wave radar to extract natural features. In *IEEE International Conference on Robotics and Automation*: IEEE.

Cox, Ingemar J. (1993). A review of statistical data association techniques for motion correspondence. *International Journal of Computer Vision* 10(1):53-66.

Davison, A.J. and D.W. Murray. (2002). Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7):865 - 880

Davison, Andrew J. (1998). Mobile Robot Navigation Using Active Vision. *Thesis*: University of Oxford.

Davison, Andrew J., Yolanda Gonzalez Cid and Nobuyuki Kita. (2004). Real-Time 3D Slam with Wide-Angle Vision. In *IFAC Symposium on Intelligent Autonomous Vehicles*. Lisbon.

Dissanayake, M.W.M.Gamini, Paul Newman, Steven Clark, Hugh F. Durrant-Whyte and M. Csorba. (2001). A Solution to the Simultaneous Localization and Map Building (SLAM) Problem. *IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION* 17(3):229-241.

Fischler, Martin A. and Robert C. Bolles. (1981). Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24(6):381 - 395.

Guivant, Jose E. (2002). Efficient Simultaneous Localization and Mapping in Large Environments. *Thesis*. Sydney: University of Sydney.

Guivant, Jose and Eduardo Nebot. (2002). *Simultaneous Localization and Map Building: Test case for Outdoor Applications*. Sydney: Australian Center for Field Robotics, Mechanical and Mechatronic Engineering, The University of Sydney.

Guivant, Jose, Juan Nieto, Favio Masson and Eduardo Nebot. (2003). Navigation and Mapping in Large Unstructured Environments. *International Journal of Robotics Research* 23(4/5): 449-472.

Herath, D. C., K. R. S. Kodagoda and Gamini Dissanayake. (2006a). Modeling Errors in Small Baseline Stereo for SLAM. In *The 9 th International Conference on Control, Automation, Robotics and Vision* (ICARCV 2006). Singapore.

Herath, D.C., Sarath Kodagoda and G. Dissanayake. (2006b). Simultaneous Localisation and Mapping: A Stereo Vision Based Approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS 2006). Beijing, China: IEEE.

Huang, Shoudong and Gamini Dissanayake. (2006). Convergence Analysis for Extended Kalman Filter based SLAM. In *IEEE International Conference on Robotics and Automation* (ICRA 2006). Orlando, Florida.

Julier, S. J. and J. K. Uhlmann. (2001). A counter example to the theory of simultaneous localization and map building. In *IEEE International Conference on Robotics and Automation*, ICRA 2001.

Julier, S. J. and J. K. Uhlmann. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE* 92(3):401-422.

Jung, I.K. (2004). Simultaneous localization and mapping in 3D environments with stereovision. *Thesis*. Toulouse: Institut National Polytechnique.

Jung, Il-Kyun and Simon Lacroix. (2003). High resolution terrain mapping using low altitude aerial stereo imagery. In *Ninth IEEE International Conference on Computer Vision* (ICCV'03).

Konolige, Kurt. (1997). Small Vision Systems: Hardware and Implementation. In *Eighth International Symposium on Robotics Research*.

Kwok, N. M. and G. Dissanayake. (2003). Bearing-only SLAM in Indoor Environments Using a Modified Particle Filter. In *Australasian Conference on Robotics & Automation*, eds. Jonathan Roberts and Gordon Wyeth. Brisbane: The Australian Robotics and Automation Association Inc.

Kwok, N. M. and G. Dissanayake. (2004). An efficient multiple hypothesis filter for bearing-only SLAM. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS 2004)

Kwok, N. M., G. Dissanayake and Q. P. Ha. (2005). Bearing-only SLAM Using a SPRT Based Gaussian Sum Filter. In I*EEE International Conference on Robotics and Automation*. ICRA 2005.

Se, Stephen, David Lowe and Jim Little. (2002). Mobile Robot Localization And Mapping with Uncertainty using Scale-Invariant Visual Landmarks. *International Journal of Robotic Research* 21(8).

Shi, Jianbo and Carlo Tomasi. (1994). Good Features toTrack. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR '94) Seattle: IEEE.

Sibley, G., G. Sukhatme and L. Matthies. (2006). The Iterated Sigma Point Filter with Applications to Long Range Stereo. In *Robotics: Science and Systems* II. Cambridge, USA.

Sibley, Gabe, Larry Matthies and Gaurav Sukhatme. (2005). Bias Reduction and Filter Convergence for Long Range Stereo. In *12th International Symposium of Robotics Research* (ISRR 2005). San Francisco, CA, USA.

Wang, Chieh-Chih. (2004). Simultaneous Localization, Mapping and Moving Object Tracking. *Thesis*. Pittsburgh, PA 15213: Carnegie Mellon University.

Williams, Stefan Bernard. (2001). Efficient Solutions to Autonomous Mapping and Navigation Problems. *Thesis*. Sydney: The University of Sydney.